A Frame Based Spoken Dialog System for Home Care

Daniele Falavigna(+), Toni Giorgino(+), Roberto Gretter(+)

(+) ITC-irst, Povo, Trento, Italy
(*) Laboratory for Medical Informatics, University of Pavia, Italy
falavi@itc.it, toni.giorgino@unipv.it, gretter@itc.it

Abstract

This paper deals with the evaluation of a dialog system, developed in order to monitor the health status of hypertensive patients, taking into account their needs, preferences and the time course of her/his disease. Statistics and dialog refinements are described with reference to data acquired, from patients, actually using the system, over a controlled clinical trial that lasted 12 months and collected 541 dialogs.

1. Introduction

The work reported in this paper describes the various development phases of a “frame based” spoken dialog system largely used for monitoring the health status of chronic patients. The system, previously described in [1, 2], allows patients affected by hypertension diseases to enter their personal clinical data, e.g., blood pressure, cardiac frequency, etc. into a database by periodically calling a telephone service and by engaging a dialog with it. The dialog takes into account the clinical histories of patients and, if needed, can give advices to them or can alert doctors. The service for handling chronic patients has arisen from the framework afforded within the EU project “Homey”1; it has been specifically developed for the Italian language using the Automatic Speech Recognition (ASR) system of ITC-irst [3, 4].

Although some recent papers [5, 6] propose to adopt formal approaches, based on probability theory [7], for modeling dialog, there is no evidence about their effectiveness to develop real telephone applications. Instead, for the tasks pursued in this work we found effective to use the approach proposed in [8, 4, 9, 10], since it allows a large degree of freedom in “manual tuning” of the system parameters, particularly for refining ASR grammars.

An overall evaluation of the health care service has been carried out on a database collected through a large number of real patients interactions. These patients, as well as their physicians, have been involved in the project by two Italian Hospitals: the Hospital of Pavia and the Hospital of Florence. In this way it was possible to derive significant statistics related to both speech recognition and dialog performance, and to the “clinical effectiveness” of the service.

Main results show that speech recognition performance, measured both in terms of Word Accuracy (WA) and Semantic Accuracy (SA), can be significantly improved after some refinements of system parameters, at the cost of manually transcribing a certain amount of recorded sentences.

This paper is organized as follows. A brief description of the whole system is given in Section 2, the field data collection

1This work has been partially funded by EU under the Fifth Framework IST project “HOMEY”: number IST-2001-32434.

2. System description

The whole system is composed of the following main functional blocks: (1) an Electronic Health Record (EHR), which contains personal data of each patient; (2) a spoken dialog interface, used by patients to enter their data in the EHR; and (3) a conventional graphical system-physician interface, to allow the latter to access patient data. Physicians and domain experts have been involved both in the design of the EHR and in the organization and refinement of dialog structure and steps.

2.1. The Electronic Health Record (EHR)

The EHR provides the dialog system both with the clinical histories of patients and with a generic domain knowledge. The dialog can benefit from these two knowledge sources: examples of how efficiently using them can be found in [11, 12]. The physician may use a conventional graphical interface to store and update patient information, while the patient can enter her/his self-measured data by means of a telephone.

2.2. The Spoken Dialog Interface

Patients periodically call a dedicated telephone number and engage a dialog with the system, which talks and interacts with them to acquire clinical data, monitors their style of life and asks about the occurrence of possible side effects. The EHR holds several values that will affect the next dialog sessions: e.g., whether the patient has been prescribed a diet, his current weight, the date of the next visit, etc.

As mentioned above the spoken dialog interface is “frame based”: it can handle mixed-initiative interactions through a dialog engine that acquires basic information (e.g. patient personal code, systolic and diastolic blood pressure, patient weight, etc) related to the specific application domain. Each basic piece of information has associated voice prompts and corresponding ASR grammars, and is activated according to a “dialog strategy” (see [4, 9] for the details). All of these components are manually defined into a document (the “dialog description”) which is fed to a dialog engine.

The system changes its run-time behavior depending on both the progress of the current call, and the clinical history of the caller. This is achieved by preserving a representation of call data, named the state vector. When the call ends, the values of some output dimensions of the state vector reflect what the user has uttered during the call. When a new call is set up for a certain patient, these values, together with those contained in the patient EHR, are processed to define the state vector for the interaction. In this way, an adaptive dialog system is realized.
2.3. Dialog issues

A preliminary evaluation was conducted over an initial set of dialogs and addressed the following issues.

- **Grammar coverage.** i.e. the percentage of sentences contained in the ASR grammars, was enhanced. Grammar coverage can be partially improved following a careful analysis of recorded and transcribed sentences.

- **Speech recognition performance.** was low. We observed several substitution errors between the Italian numbers “60” and “70”, even inside composite numbers such as “164”. This is mainly due to the fact that the two words are acoustically similar, as they differ only in the SAMPA phones /s/ in “60” which becomes /t/ in “70”. In some cases, patients spontaneously tried to correct data by spelling digits, e.g. “67, 6 7” (“sixty-seven, six seven”).

- **Speech synthesis.** suffered of the same problem. We observed that sometimes the patient did not confirm a value because she/he misunderstood the confirmation prompt (e.g. “diastolic blood pressure 76, do you confirm?”).

- **Multiple confirmations.** of several clinical values collected during the call was not effective. Many patients found difficult to correctly answer to the question “Tell me which data is wrong”. Moreover, in case of multiple errors, they sometimes corrected only one of them, leaving the others unchanged.

- **Redundant confirmation questions.** were removed from rarely used sub-dialogs.

2.4. Dialog refinements

To overcome the problems described above, the following modifications were applied to the dialog:

- minor modifications to both speech recognition grammars and dialog strategy; this has allowed to improve language coverage and to avoid unnecessary confirmations.

- grammars for recognizing numbers were modified to allow for the “spelled-digit” mode, in different combinations. This means that each value (e.g. 123) can be pronounced equivalently as:
  
  (number) (e.g. 123);
  
  (digit+) (e.g. 1 2 3);
  
  (number digit+) (e.g. 123 1 2 3);
  
  (digit+ number) (e.g. 1 2 3 123).

This modification leads to a better grammar coverage, and make it easier to correct ASR errors.

- **Multiple confirmations** was replaced by immediate explicit confirmation.

Patients were informed of the updates via the system’s welcome prompt.

3. Data collection

The evaluation of the Chronic Disease demonstrator was conducted in two steps. First an internal trial, carried out with the help of volunteers, has addressed the main technical and usability issues. A thorough clinical trial was then designed to assess issues involved in the introduction of the web-based clinical record, augmented by the automated adaptive dialog system. The data collection architecture keeps into account two sources of information: (1) call logs, and (2) clinical information about the patients. The former is gathered from information generated, at run time, by the computer telephony platform.

3.1. Clinical trial design

For the clinical trial, started in August 2003, a group of 125 patients was enrolled and split into two groups (control and treatment). Patients had not been specifically trained for using the telephone system. The clinical trial was designed as a two-arm controlled study. After an initial visit, patients in the treatment group were given a toll-free telephone number to call, and a randomly generated 6-digit personal access code. An histogram with patients’ age is reported in the following:

<table>
<thead>
<tr>
<th>age range</th>
<th>3</th>
<th>9</th>
<th>33</th>
<th>50</th>
<th>29</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≤ 29</td>
<td>30-39</td>
<td>40-49</td>
<td>50-59</td>
<td>60-69</td>
<td>≥ 70</td>
</tr>
</tbody>
</table>

All of the data on which this review is based come from actual clinical trial calls. Overall, 541 calls (dialogs) were collected in the time span considered for this review. Parallel to this, 301 records were entered by physicians during face-to-face visits.

3.2. Speech data

As seen above, all of the telephone calls were recorded. Each of them corresponds to a patient dialog, where the patient can be identified after the verification of a 6-digit personal number. Dialogs were divided into 3 sets: test patients, performed by researchers which used their own (dummy) codes; anonymous patients, when the system failed to recognize a valid code; true patients (further subdivided into males and females), performed by chronic patients correctly identified by their code.

All of the dialogs of true patients until September 2004 have been orthographically transcribed. Three data sets have been defined:

- **feb04:** dialogs collected between August 2003 and February 2004. In total they amount to 5485 speech utterances, divided as given in Table 1;

- **aug04:** dialogs collected after the dialog refinements (end May 2004) between June 2004 and August 2004. They contain 2277 sentences, and were mainly used to verify the effectiveness of the modifications described above.

- **sep04:** dialogs collected between August 2003 and September 2004. They include both feb04 and aug04 data and contain 9527 speech utterances, for a total of 541 dialogs, 54 speakers as shown in Table 1.

4. Evaluation

Evaluation results collected from the clinical trial of the hypertension management prototype have been analyzed along the following major directions: (a) verification of technical aspects of the chosen ASR technology, grammars, and telephone interface; (b) assessment of effects of the combined system (web-

<table>
<thead>
<tr>
<th>Patients Feb04</th>
<th>Speakers</th>
<th>Utterances</th>
<th>Dialogs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>11</td>
<td>1494</td>
<td>113</td>
</tr>
<tr>
<td>Anonymous</td>
<td>20</td>
<td>207</td>
<td>29</td>
</tr>
<tr>
<td>True female (feb04-f)</td>
<td>19</td>
<td>1859</td>
<td>118</td>
</tr>
<tr>
<td>True male (feb04-m)</td>
<td>25</td>
<td>1925</td>
<td>117</td>
</tr>
<tr>
<td>True (feb04)</td>
<td>44</td>
<td>3784</td>
<td>235</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patients Sep04</th>
<th>Speakers</th>
<th>Utterances</th>
<th>Dialogs</th>
</tr>
</thead>
<tbody>
<tr>
<td>True female (sep04-f)</td>
<td>21</td>
<td>4242</td>
<td>244</td>
</tr>
<tr>
<td>True male (sep04-m)</td>
<td>33</td>
<td>5285</td>
<td>297</td>
</tr>
<tr>
<td>True (sep04)</td>
<td>54</td>
<td>9527</td>
<td>541</td>
</tr>
</tbody>
</table>
accessible database plus home monitoring service) on patients clinical state and perception of selves health.

Despite the care put into the design of the dialog application, questions and grammars, the fact that untrained users interact with such a complex system poses remarkable usability challenges.

4.1. Call diagrams

A tool both to monitor system effectiveness and to rapidly discover potential problems has been developed.

Figure 1: Overview of the usage patterns of the telephone system. Enlarged boxes show patterns corresponding to different situations.

Figure 1 provides one view of the usage patterns of the telephone system, over time, for a subset of the treatment patients enrolled into the clinical trial. Each patient is shown as a Cartesian plane in a separate box. A dot is placed on the plane for each call; the x coordinate relates to the day the call was received (day zero stays for the begin of the trial, right boundary stays for actual time). The y axis displays the cumulative number of calls. Points have a different shape to indicate whether the dialog reached the final acknowledgment message (triangle) or not (small circle). This representation was chosen to distinguish conversations that ended correctly or not; however, often patients hang-up before the final message, but after having confirmed the more important clinical data.

Patterns in these plots can be recognized by a simple inspection. Call diagrams like (a) correspond to people that regularly call the system and interact successfully with it. They show dominantly “good-end” calls. The slope of curve (a) shows that the respective patient followed the physicians direction to call more frequently, once a week, during the first two months of the systems use, then to reduce the number of calls to every other week. People which call the system regularly, but experience usage problems, may show up as in diagram (b), where there are a large fraction of “circle” symbols, but the curve does not have remarkable jumps: this means that on the average the call frequency was kept as required, despite of the outcome of specific calls. Patients that regrettably stop using the system altogether are also quite easily distinguished, as in (c), where the last call is far from the right boundary. It is worth noting that not all quitters have a record of unsuccessful previous calls.

4.2. Effect on Patient clinical status

Another type of evaluation concerns whether the patient’s health benefits from a regular check, like the one provided by an automatic system. We chose to evaluate the clinical outcomes according to the 24-hour average blood pressure (ABPM). ABPM samples patients blood pressure at regular time intervals over 24 hours (including day and night). Patients approximately undergo ABPM every 6 months. At the time this report was prepared, only a limited fraction of the patients has undergone at least two ABPM measurements. Taking control and treatment patients together, there is an hint ($p < 0.10$) that the 24h average systolic blood pressure decreases. However, due to the limited sample size, further investigation with more data is necessary to derive statistically significant results.

4.3. Dialog refinements

Some weeks after the refinements above mentioned, a subjective evaluation of the new dialogs was performed to verify their effectiveness. In particular we observed that:

- the “normal” (i.e., explicit) confirmation strategy worked better than the multiple (combined) confirmation;
- dialogs looked more fluent, due to some minor changes;
- spelled digits were particularly appreciated. At the beginning, most of the patients tried this new function when entering numbers, and after a while they used it only for corrections. Considering that numbers can now be entered as either normal numbers, e.g. sixty-seven (number), or spelled-out, e.g. six seven (digit+), in the new dialogs the former (number) was used 455 times, the latter (digit+) was used 116 times. Statistics measured over the semantic form chosen by callers are given below, where (yes-no) means that the sentence was uttered during a confirmation.

- as will be seen below, both coverage and recognition rate improved significantly.

Furthermore, after the dialog corrections, the fraction of dialogs for which the telephone was hung-up before final salutation dropped remarkably.

4.4. Grammar Coverage

Grammar coverage has been directly evaluated by summing the coverage of each grammar used in the dialog description. The number of different grammars was 43 before the modifications, 36 after them. For each grammar, two recognition lists can be defined: the first one contains the audio files recorded in a certain dialog state, the second contains the subset of sentences covered by the grammar active in the given dialog state. Improving coverage normally leads to a trade-off: better recognition rate on the whole list, worse recognition rate on the covered subset. By summing up the contributions for each grammar the following results, expressed as the percentage of sentences that can be generated by the set of ASR grammars used in the system, were obtained.

<table>
<thead>
<tr>
<th>data set</th>
<th>coverage</th>
<th>sent. covered/tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>feb04</td>
<td>90.46%</td>
<td>3423 / 3784</td>
</tr>
<tr>
<td>aug04</td>
<td>95.87%</td>
<td>2183 / 2277</td>
</tr>
<tr>
<td>set04</td>
<td>94.18%</td>
<td>8722 / 9261</td>
</tr>
</tbody>
</table>

Note that for the set04 set we used the last version of the grammars, thus only the sentences recognized with grammars survived after the refinement step were considered (9261 sentences out of 9527). From this table it can be seen that the percentage of out-of-grammar sentences was significantly reduced after refinement.
4.5. Word Accuracy

Table 2 reports results for the various data sets defined in Section 3.2, as well as results for the two corresponding gender dependent subsets. In the table, StrRR stands for String Recognition Rate (percentage of correctly recognized sentences), UnitRR stands for Unit Recognition Rate (percentage of correctly recognized units), Units is the number of units pronounced in the data set, Errs reports the number of Deletion, Insertion, Substitution errors. For this evaluation units correspond to words.

<table>
<thead>
<tr>
<th>data set</th>
<th>StrRR</th>
<th>UnitRR</th>
<th>#Units</th>
<th>#Errs (D+I+S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>feb04</td>
<td>71.85%</td>
<td>72.77%</td>
<td>6218</td>
<td>1693 (341+394+955)</td>
</tr>
<tr>
<td>feb04-cov</td>
<td>85.25%</td>
<td>87.00%</td>
<td>5195</td>
<td>738 (88+204+446)</td>
</tr>
<tr>
<td>feb04-f</td>
<td>73.00%</td>
<td>77.00%</td>
<td>3193</td>
<td>1014 (229+167+618)</td>
</tr>
<tr>
<td>feb04-m</td>
<td>82.55%</td>
<td>77.55%</td>
<td>3025</td>
<td>679 (147+165+367)</td>
</tr>
<tr>
<td>aug04</td>
<td>88.98%</td>
<td>72.77%</td>
<td>3488</td>
<td>390 (164+56+170)</td>
</tr>
<tr>
<td>aug04-cov</td>
<td>88.82%</td>
<td>91.20%</td>
<td>3287</td>
<td>253 (85+40+128)</td>
</tr>
<tr>
<td>sep04</td>
<td>85.78%</td>
<td>77.00%</td>
<td>13434</td>
<td>2156 (879+317+960)</td>
</tr>
<tr>
<td>sep04-cov</td>
<td>90.09%</td>
<td>72.77%</td>
<td>12908</td>
<td>1090 (271+172+647)</td>
</tr>
<tr>
<td>sep04-f</td>
<td>84.33%</td>
<td>91.20%</td>
<td>6726</td>
<td>1106 (407+162+537)</td>
</tr>
<tr>
<td>sep04-m</td>
<td>87.00%</td>
<td>86.22%</td>
<td>7618</td>
<td>1050 (472+155+423)</td>
</tr>
</tbody>
</table>

Note that female patients (feb04-f, sep04-f) perform significantly worse that male ones (feb04-m, sep04-m), even if the gap in the sep04 set is greatly reduced with respect to the feb set. Note also that the dialog refinement not only led to a better grammar coverage, as previously discussed, but even to an increase in UnitRR. This comparison is not straightforward, since the data sets are different. Moreover, other causes (for instance, the expertise level of the patients) may affect the overall results.

In Figure 2, dialogs have been grouped to highlight how user expertise impacts on recognition rate. The first point on the x-axis groups each first and second dialog of each patient; the second point groups dialogs from number 3 to number 4, and so on. The size of the data decreases as the dialog number increases (only a few patients reached dialog number 30); nevertheless there is a clear improvement in recognition rate as user expertise grows.

Figure 2: Recognition rate as a function of user expertise. Note that the size of the data reduces significantly after 20 dialogs.

4.6. Semantic Accuracy

In a dialog application it makes sense to compute semantic accuracy, which represent the dialog behavior better than word accuracy. Semantic accuracy is reported below. Note that StrRR, UnitRR etc. have the same meaning as before, except that units are now semantic concepts rather than words.

5. Conclusions

We have evaluated the field performance of a mixed initiative spoken dialog system aimed at offering health care assistance to hypertensive patients. The system has been demonstrated effective after some manual refinements of both speech recognition grammars and dialog strategy. Future works will address the task of using additional sources of knowledge (medical, clinical, etc) to improve the clinical effectiveness of the whole system.

We thank Ivano Azzini for the development of the first version of the dialog description.

6. References